
Journal of Urban Health: Bulletin of the New York Academy of Medicine Vol. 80, No. 2, Supplement 1 2003

Syndromic Surveillance Using Automated Collection of Computerized Discharge Diagnoses

William B. Lober, Lisa J. Trigg, Bryant T. Karras, David Bliss, Jack Ciliberti, Laurie Stewart, and Jeffrey S. Duchin

ABSTRACT *The Syndromic Surveillance Information Collection (SSIC) system aims to facilitate early detection of bioterrorism attacks (with such agents as anthrax, brucellosis, plague, Q fever, tularemia, smallpox, viral encephalitides, hemorrhagic fever, botulism toxins, staphylococcal enterotoxin B, etc.) and early detection of naturally occurring disease outbreaks, including large foodborne disease outbreaks, emerging infections, and pandemic influenza. This is accomplished using automated data collection of visit-level discharge diagnoses from heterogeneous clinical information systems, integrating those data into a common XML (Extensible Markup Language) form, and monitoring the results to detect unusual patterns of illness in the population. The system, operational since January 2001, collects, integrates, and displays data from three emergency department and urgent care (ED/UC) departments and nine primary care clinics by automatically mining data from the information systems of those facilities. With continued development, this system will constitute the foundation of a population-based surveillance system that will facilitate targeted investigation of clinical syndromes under surveillance and allow early detection of unusual clusters of illness compatible with bioterrorism or disease outbreaks.*

KEYWORDS *Biological warfare, Bioterrorism, Data collection, Database, Informatics, Information systems, Sentinel surveillance.*

INTRODUCTION

Syndromic Surveillance

Bioterrorism has become a threat to national security in the 21st century. Vulnerabilities have been discovered in the security of bioterrorism agents developed by weapons programs of both the former Soviet Union and the United States. The recent illnesses and deaths attributed to *Bacillus anthracis* have further emphasized the need for the capability to detect and respond to biological weapons attacks. Even prior to the terrorist events of 2001, previous bioterrorist attacks led to growing concern about US public health vulnerability to such attacks.

Experience with a manual, clinician-based (versus automated, information system–based) emergency department (ED) surveillance system employed during the

Dr. Lober, Ms. Trigg, Dr. Karras, and Mr. Bliss are with the Clinical Informatics Research Group, University of Washington, Seattle; Ms. Stewart and Dr. Duchin are with Public Health—Seattle and King County, Seattle, Washington; and Dr. Ciliberti is with the Overlake Hospital Medical Center, Bellevue, Washington.

Correspondence: Bill Lober, MD, Research Assistant Professor, University of Washington, MS 357240, 1959 Northeast Pacific Street, Seattle, WA 98195. (E-mail: lober@u.washington.edu)

1999 World Trade Organization (WTO) Ministerial in Seattle, Washington, suggested that syndromic surveillance is a potentially useful means of monitoring selected clinical illnesses in the ED setting.[1] The Seattle WTO surveillance project identified the need for the development and evaluation of active surveillance systems to provide early detection of bioterrorist attacks at the local government level. However, prior to 2001, there was little in the literature on the design, development, and implementation of surveillance systems to meet this need.[2,3] During the fall 2001 American Medical Informatics Association meeting, a roundtable meeting was held for those engaged in developing or interested in information system–based surveillance systems. The report from that meeting summarizes eight of these surveillance systems.[4]

Objectives

An optimal response to bioterrorist attacks is predicated on the reliability of early detection methods. The goal of the SSIC project is to develop and support an automated, nationwide surveillance system that will facilitate early detection of bioterrorist attacks. Specifically, through SSIC, we (1) employ automated collection of data from heterogeneous source systems; (2) normalize clinical syndromic data and store it in a centralized database; (3) provide secure, remote access to this data for public health staff using aberration detection software; (4) characterize baseline frequencies of certain diagnoses and diagnostic clusters as seen through the surveillance system; and (5) are exploring strategies for further processing of data to enhance event detection.

Project Overview

The Clinical Informatics Research Group (CIRG) at the University of Washington and Public Health—Seattle and King County (PHSKC) have collaborated to develop an automated system that collects data on the presenting complaints and discharge diagnoses of ED and primary care patients. The pilot project was done with the cooperation of three EDs at unrelated hospitals, as well as at a university-based system of primary care clinics, and has been operational since June 2001. The data are used operationally in daily surveillance by PHSKC epidemiologists, who run aberration detection software on the data each morning. The output of this software is evaluated, in combination with input from other reporting systems (such as emergency medical services dispatches, school absenteeism reports, and reports of unexplained deaths) by the epidemiologists, who use all sources to determine whether to begin an investigation.

While they are available, these data are presently of unproven value in detecting an outbreak of disease. We have concerns about the limited types of data we collect, about the difficulty of comparing these types across institutions, and about data quality problems introduced by our present strategy of deidentified data collection. For these reasons, we have not done extensive analysis of the data, concentrating instead on improving infrastructure and techniques for data collection.

METHODS

Sentinel Events

Several candidate agents for potential bioterrorism attacks are characterized by aerosol dispersion that results in acquisition of infection by inhalation. These infec-

tions may present as respiratory syndromes or influenzalike illnesses. We monitor data on patient visits to the participating EDs and primary care clinics for the occurrence of either sentinel *International Classification of Diseases, 9th Revision* (*ICD-9*) diagnoses (Table 1) or terms identified by keyword searches of chief complaint fields. Patient records identified by either mechanism comprise the sentinel surveillance events. The *IDC-9* codes and keywords currently monitored were selected through an expert review process and are based on their likelihood of identifying clusters of patients with influenzalike illness or other syndromes compatible with disease due to agents of biological warfare. The selection of codes was modified by the specific *ICD-9* and free-text options offered to clinicians by the discharge diagnostic software in use at each site.

System Overview

The Syndromic Surveillance Information Collection (see Fig. 1) system currently collects data on sentinel events from four clinical information systems. For each individual site, we tailor data extraction and transmission software to the specific requirements of the information system and security environment and install it so that it queries its host periodically. These data are transmitted to a centralized cluster of servers, where they are normalized to a common format, represented by the XML schema in Fig. 2, and then stored in a relational database. These uniform, multisite data are then made available for secure queries from specific PHSKC workstations, on which aberration detection software is run. This process is described in more detail in the subsequent sections of this article.

Reporting Sites

The surveillance system presently encompasses four health care systems in King County, Washington, on either side of Lake Washington, as shown in Fig. 3. While they represent a convenience sample for our initial system development, they also represent good geographic dispersion, wide patient catchment areas, and diverse patient populations, including children and adults.

ED A is a tertiary care community hospital site with a mixed adult and pediatric practice serving approximately 52,000 patients per year. ED B is a university

TABLE 1. Subset of *ICD-9* codes and key words currently used

Diagnosis or free text	*ICD-9*
Viral syndrome, pediatric	079.9
Pneumonia, viral	480.9
Influenza	487.1
Flu	
Enterocolitis	009.0
Diarrhea, infectious	009.2
Viral meningitis	047.9
Bronchitis	466.0
Pneumonia, pediatric	486
Diarrhea, pediatric	787.91
Measles	055.9
Pleurisy	511.0

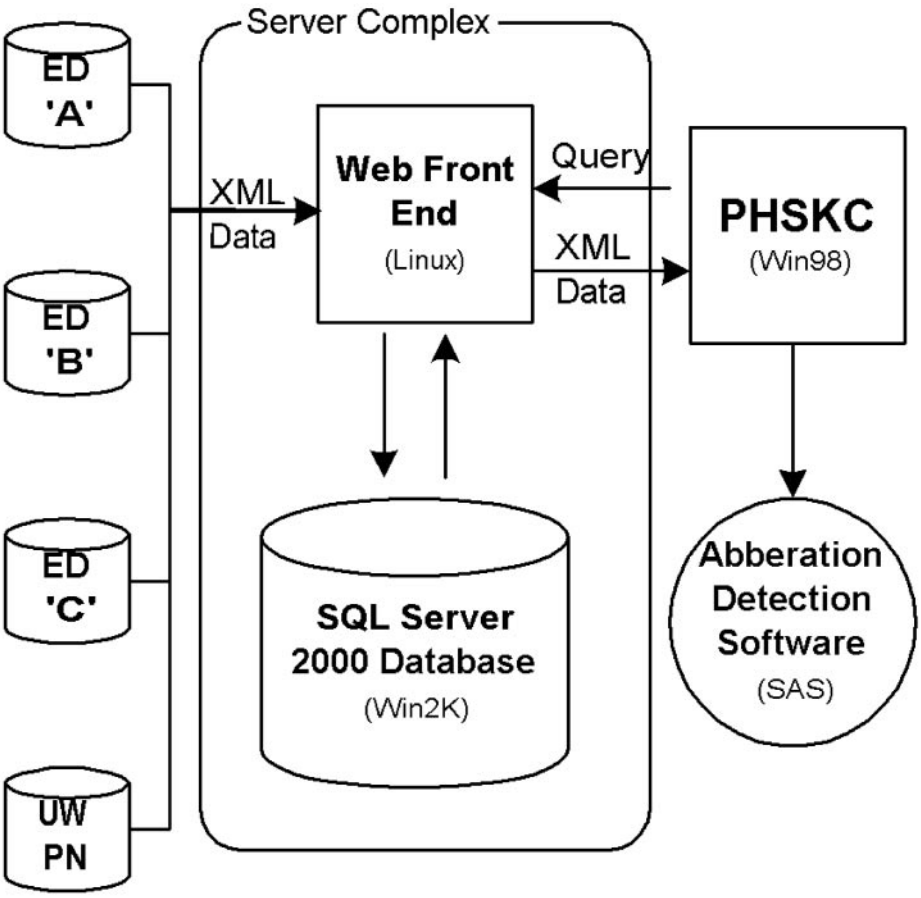

FIGURE 1. System overview data are collected and stored and can be queried centrally.

hospital with a primarily adult practice. Site B includes an urgent care facility, which has data that are managed by the same system. Together, these sites represent approximately 30,000 patient visits per year. ED C is a regional pediatric tertiary care center and referral center with an annual volume of approximately 26,000 patients. The Primary Care Network (stars in Fig. 3) is a university-affiliated clinic serving a mixed population of patients from the city and surrounding communities. This network serves approximately 240,000 primary care patients per year.

Data Collection

The three participating EDs have implemented clinical database management systems based on Spacelabs' ED Chart products (Spacelabs Medical, Redmond, WA). Although the same product is implemented at the three sites, the hospitals are independent and have different software configurations, networks, security infrastructures, information technology (IT) departments, and administrations. ED A was the host of much of the development work on the ED Chart product and has significant expertise with the software. ED B and ED C implemented ED Chart 2 to 3 years ago, upgrading from an earlier, Macintosh-based, precursor (Orca Systems, Belle-

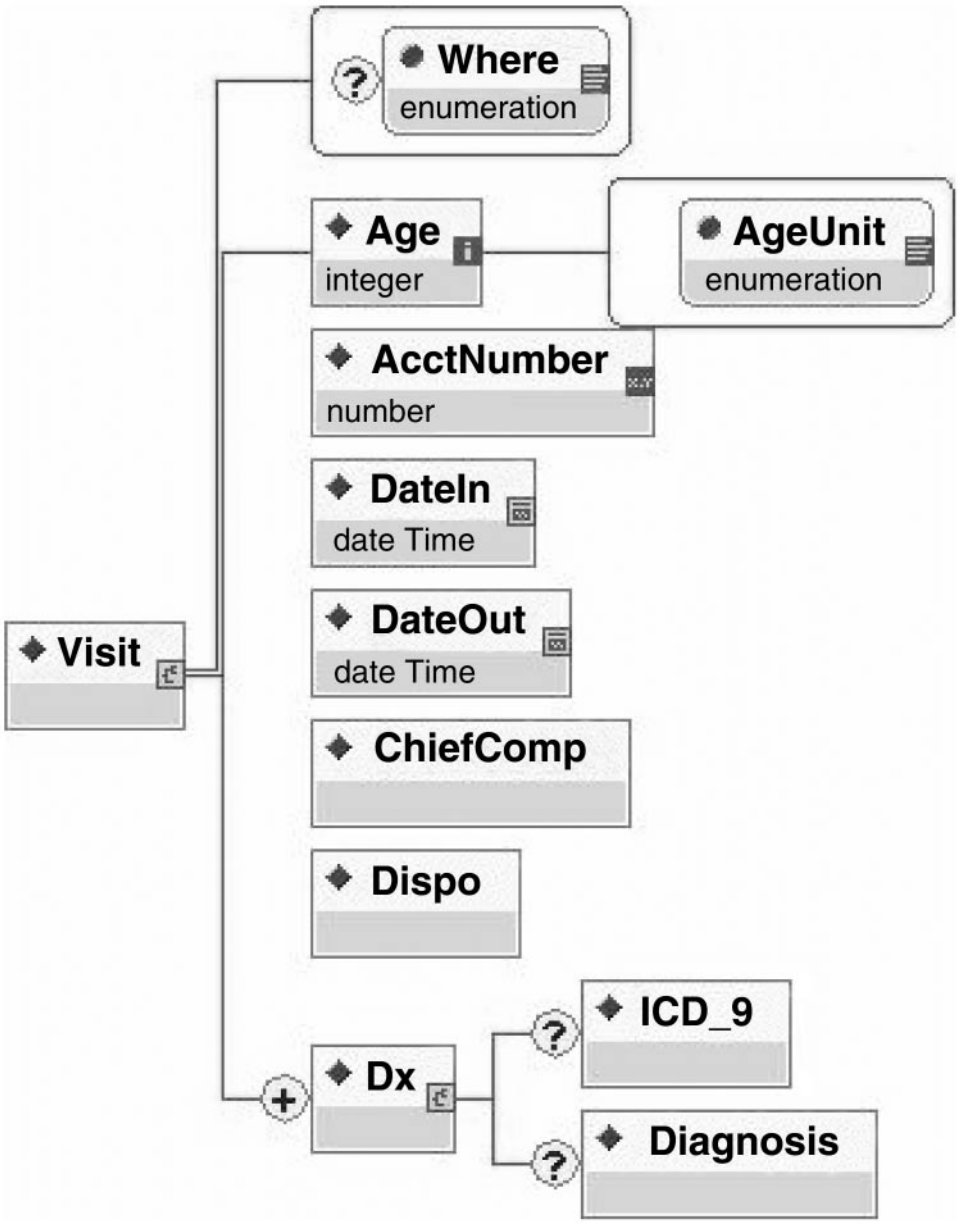

FIGURE 2. Graphic representation of Extensible Markup Language (XML) schema, organized hierarchically at the visit level. Data are normalized to this schema on the central server.

vue, WA). At ED A, the data collection software is installed on the server, while at the other two sites, it is installed on a client workstation.

ED Chart is a FoxPro (Microsoft Corp., Redmond, WA) application that uses a client-server model. The vendor does not publish the details of its database schema. However, our query software is highly customizable, and we have been able to extract data to satisfy our schema. The data extraction software for these three systems is written in FoxPro, which supports XML data representation. In our

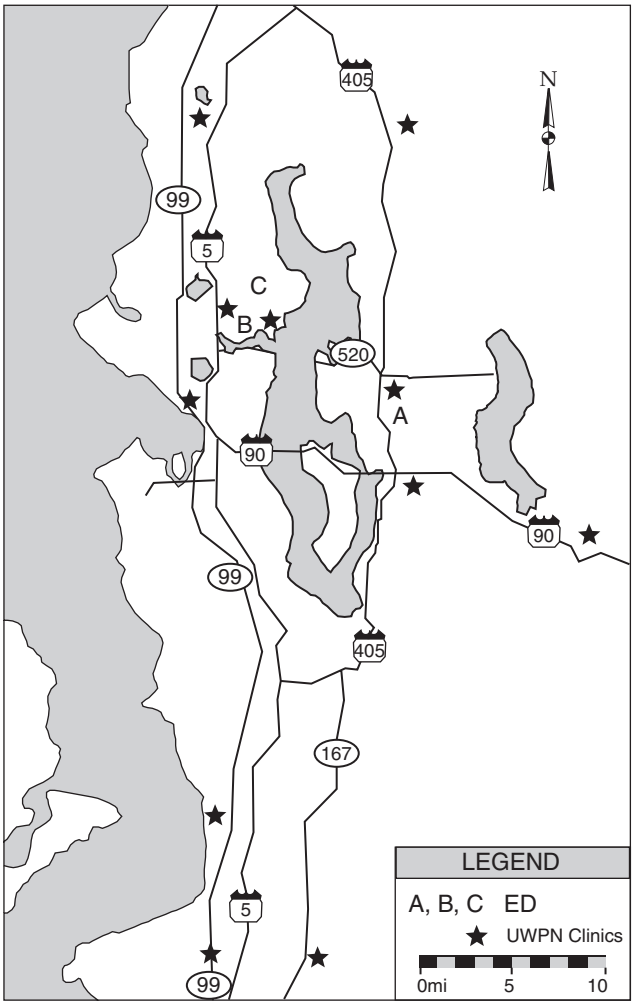

FIGURE 3. Sentinel site distribution, showing emergency departments and clinics reporting data as of September 2002.

system, the FoxPro application is paired with an open source package to provide secure communications (cURL, http://curl.haxx.se). The cURL program is a "command line URL" program that supports Secure Sockets Layer (SSL) encryption and allows both automated and application program–driven, forms-based submission of data, including file uploads, using the secure Hypertext Transfer Protocol (HTTP).

The participating primary care network uses a different information system architecture. This network consists of nine clinics distributed throughout the region that share the same clinical information system (EpicCare, Epic Systems Corp., Madison WI). This vendor-supplied system uses Crystal Reports as its report-writing package. Crystal Reports is a flexible package, and the clinic network IT staff has good internal development capabilities for the customization of reports. These reports, locally reformatted prior to secure transmission, serve as the on-site, data-gathering tool at the primary care network.

Central Servers

The SSIC Database Server consists of three components: (1) a data collection application to receive the secure transmissions from the heterogeneous clients, perform a final data conversion, and store the data; (2) a database management system to support data storage and queries; and (3) a secure, Web-based query processor. The XML schema shown in Fig. 2 also plays a role at the database level as it is expressed in a format (XDR Schema) that allows us to generate database tables automatically and to import bulk data. The database is implemented using Microsoft SQL Server 2000 with Web Release 1 extensions. The three components of the database server are split across two physical machines as depicted in Fig. 1. The data collection application, which is written in Perl (http://perl.com), and the query processor, which is written in PHP (http://php.net), both run on a Linux server (Slackware 7.1, http://slackware.com). The database runs on a Windows 2000 server, which is only allowed to communicate with the Linux front-end computer. We believe this maximizes functionality and security.

Real-Time Queries/Automated Data Delivery

The Linux data collection application processes real-time queries into the SQL Server database using secure forms-based submission of constrained query parameters, which are translated into SQL on the Windows machine. The resulting row sets are delivered via a Hypertext Transfer Protocol transaction over a Secure Sockets Layer connection. We have implemented automated data delivery using the same mechanism. In this case, the workstation requesting data sends a scheduled query request, using cURL to ensure secure communications, and stores the resulting row set locally for further processing. This mechanism is easily integrated with local database and statistical applications.

Security Considerations

Despite the absence of patient identifiers, we treat all data sent from any client system as if patient identified. This means that we transmit data only over secure, encrypted links using validated certificates, and we only store the data on servers that are logically and physically secure. We use a "strong" authentication system, using "hard," or token-based, certificates stored on smart cards (SchlumbergerSema, Austin, TX) to identify all clients of the database system, both data sources such as the ED information systems and data clients such as the epidemiologist end users.

RESULTS

Current Data Collection

We collect daily data from all sites and have obtained historical data. At two sites, these data extend to March 1999; the other two sites have a less-complete record. The database contains records of 51,543 patient visits that resulted in a report of a sentinel event, from all sites, though February 2002.

The development group has performed only minimal analysis on the data, but has instead concentrated on improving the infrastructure and expanding the data collection network. The results we present here were developed primarily for the exploration of data quality, but do offer some sense of the content of the data.

Figure 4 compares the rates of occurrence of sentinel events at the four sites. These rate calculations are based on monthly event counts, but use an annualized

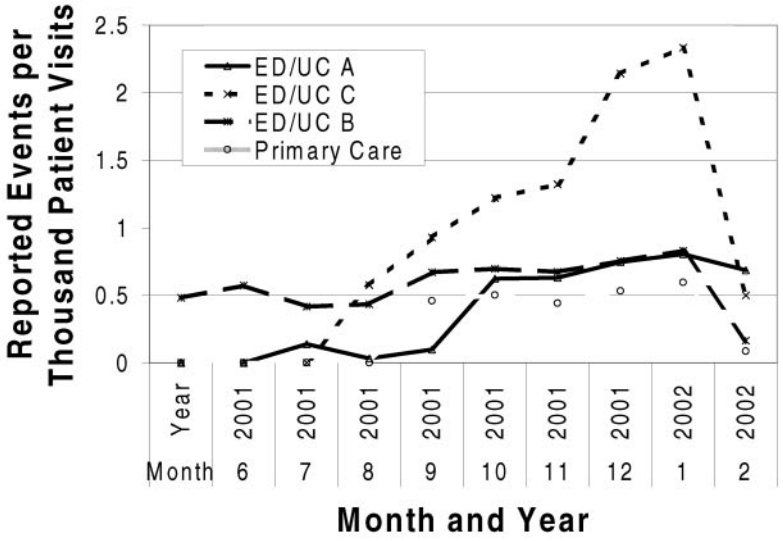

FIGURE 4. Site-specific rates, adjusted for annualized patient volume, of visits identified as sentinel events.

denominator as total monthly visits were not available for all sites. This does not adjust for seasonal differences in overall patient volumes, which may mask seasonal or episodic trends in event rates. This is clearly not as accurate as collecting precise visit numbers as a denominator, which is a weakness we intend to address. The rates from different sites roughly correlate, although the increase at the children's hospital was much more marked during this period.

Table 2 compares the incidence of two very common influenzalike sentinel events: upper respiratory tract infections (adult and pediatric combined) and pneumonia. These events are reported for each ED site and, in the aggregate, for all nine primary care sites. The periods are June 11 to September 10, 2001; September 11 to December 10, 2001; and December 11, 2001, to March 6, 2002. This is an example of the level of stratification that is easily available from the database. In the aberration detection software, these counts are compared year over year to account for seasonal variation; however, we have not yet made this comparison in our external analysis.

Confidentiality Issues

Our pilot project does not include transmission of any patient identified data, in accordance with our institutional review board approval. However, subsequent to the start of the project, Washington State notifiable disease reporting regulations were revised to include mandatory reporting by health care providers and hospitals of clusters of cases of illness compatible with bioterrorism as well as suspected cases of illness due to potential agents of bioterrorism. When a report is made of a critical condition or an increase beyond the expected number of cases is detected, an investigation is initiated by public health staff, who then contact the hospital(s) and clinician(s) caring for the patient(s). In the near future, we intend to begin transmission of all required fields for notifiable disease reporting, including complete demographic information.

TABLE 2. Sentinel event distribution stratified by site and by three periods in 2001–2002

	June–September	September–December	December–March
Upper respiratory infection, adult and pediatric (465.9)			
A	8	148	184
B	66	114	165
C	8	404	610
Primary Care	N/A	1,524	1,861
Total	82	2,190	2,820
Pneumonia (486)			
A	13	141	256
B	73	84	88
C	2	159	259
Primary Care	N/A	84	220
Total	88	468	823

N/A, not applicable.

Challenges

The greatest challenges of this project have been administrative, practice, and security issues rather than technical development. Several other hospitals have expressed willingness to participate in the data collection and syndromic surveillance efforts. However, a number of steps must be taken to incorporate a new site. These include establishing administrative contacts and agreements, creating relationships with IT groups, and clarifying data elements, architectures, security, and policy issues. All of these must precede development and implementation.

One challenge arises from the variation in coding practices. The same illness may be coded differently at different sites or by different practitioners at the same site; this might be driven by variations in education or reimbursement strategies. To address this, we plan to cluster codes in syndrome groups (J. A. Pavlin, Department of Defense Global Emerging Infections Surveillance and Response System, personal communication, November 18, 2001).

An additional challenge is to integrate guidance on security from different sources. For instance, the National Emerging Disease Surveillance System,[5] developed by the Centers for Disease Control and Prevention, and the Health Insurance Portability and Accountability Act's Standards for Security and Electronic Signatures[6] were developed to address different sets of concerns, have different priorities and constituencies, and thus apply in different circumstances.

DISCUSSION

We have made substantial progress in building an infrastructure to automate syndromic surveillance, have demonstrated a series of technologies to collect data from heterogeneous information systems, and have more than a year of experience with this deployed system. The impetus for our project was the Seattle WTO Conference, which provided both PHSKC and the local hospitals with practical experience in deploying ED-based surveillance, although using manual rather than automated

data collection. However, the present system has gone well beyond the scope and structure of manual surveillance.

At this time, we cannot evaluate the utility of this type of automated, electronic syndromic surveillance from a public health perspective because we do not yet have a large enough population base under surveillance. Following the expansion to additional sites, we plan to evaluate the system by comparing the surveillance attributes to traditional surveillance reporting methods using the recently revised criteria of the Centers for Disease Control and Prevention for evaluation of surveillance systems.[7] Ultimately, we will need to evaluate the cost-effectiveness of electronic syndromic surveillance in comparison with traditional reporting.

Our immediate plans for developing the system include (1) expanding the number of source data systems to include other EDs, urgent care settings, and primary care clinics; (2) increasing the number of specific data elements collected; and (3) achieving tighter integration with visualization and aberration detection software. In the long run, we believe that if this type of syndromic surveillance system proves to be of public health utility, then the potential to extend surveillance to include other notifiable conditions will further increase the value of the system with little additional cost.

ACKNOWLEDGEMENT

We gratefully acknowledge the support of the Centers for Disease Control and Prevention through a state bioterrorism preparedness grant (B2 section; U90/CCU017010-02); of the Washington State Department of Public Health (N11264 and N10068); and of Public Health Seattle and King County (D29174D). In addition, we would like to thank Ken More Cam for assistance in preparation of the figures.

REFERENCES

1. Plough A. WTO enhanced surveillance project—local and national collaboration leads to success. *EPI-LOG Communicable Dis Epidemiol News*. December 1999:12.
2. Waeckerle JF. Domestic preparedness for events involving weapons of mass destruction. *JAMA*. 2000;283:252–254.
3. Centers for Disease Control and Prevention. Biological and chemical terrorism: a strategic plan for preparedness and response. *MMWR Morb Mortal Wkly Rep*. 2000;49(RR4):1.
4. Lober WB, Karras BT, Wagner MM, et al. Roundtable on bioterrorism detection: information systems-based surveillance. *J Am Med Inform Assoc*. 2002;9:105–115.
5. National Electronic Disease Surveillance System (NEDSS): a standards-based approach to connect public health and clinical medicine. *J Public Health Man Pract*. 2001;7:43–50.
6. Security and Electronic Signature Standards; Proposed Rule, 45 CFR Part 142 (August 12, 1998). Available at: http://aspe.hhs.gov/admnsimp/nprm/seclist.htm. Accessed January 20, 2003.
7. Centers for Disease Control and Prevention. Updated guidelines for evaluating surveillance systems: recommendations from the guidelines working group. *MMWR Morb Mortal Wkly Rep*. 2001;50(RR13):1–35.